Supporting Information file for:
Development of Liquid Chromatographic Retention Index based on Cocamide Diethanolamine Homologous Series (C(*n)*-DEA)

Reza Aalizadeh [*,†], Varvara Nikolopoulou [a, †], Nikolaos S. Thomaidis [*,†]

[†] Laboratory of Analytical Chemistry, Department of Chemistry, National and Kapodistrian University of Athens, Panepistimiopolis Zografou, 15771, Athens, Greece

[a] These authors contributed equally

* Corresponding Authors
Reza Aalizadeh
Laboratory of Analytical Chemistry,
Department of Chemistry,
National and Kapodistrian University of Athens,
Panepistimiopolis Zografou, 157 71 Athens, Greece
Tel.:+302107274576 –Fax: +302107274750
E-mail address: raalizadeh@chem.uoa.gr

Nikolaos S. Thomaidis
Laboratory of Analytical Chemistry,
Department of Chemistry,
National and Kapodistrian University of Athens,
Panepistimiopolis Zografou, 157 71 Athens, Greece
Tel.:+302107274317 –Fax: +302107274750
E-mail address: ntho@chem.uoa.gr

Table of Content

SI 1. Materials and Method

SI 1.1. Instrumentation

In RPLC and LC 1 and 2, the chromatographic separation was performed on an Acclaim RSLC C18 column (2.1 × 100 mm, 2.2 µm) from Thermo Fisher Scientific (Driesch, Germany) preceded by a guard column, ACQUITY UPLC BEH C18 1.7 µm, VanGuard Pre-Column, Waters (Ireland), thermostated at 30 ˚C. Mobile phase composition in positive ionization mode (PI) is (A) H_2O:MeOH (90:10) with 5 mM ammonium formate and 0.01% formic acid and (B) MeOH with 5 mM ammonium formate and 0.01% formic acid. For the negative ionization mode (NI), the mobile phase is (A) H_2O:MeOH (90:10) with 5 mM ammonium acetate and (B) MeOH with 5 mM ammonium acetate. The gradient elution program was the same for the two ionization modes and the chromatogram lasts 15.5 min, with 5 min of re-equilibration of the column for the next injection. It starts with 1% B with a flow rate of 0.2 mL min^{-1} for 1 min and it increases to 39 % in 2 min (flow rate 0.2 mL min^{-1}), and then to 99.9 % (flow rate 0.4 mLmin^{-1}) in the following 11 min. Then, it keeps constant for 2 min (flow rate 0.48 mL min^{-1}) and then, initial conditions were restored within 0.1 min and the flow rate decreased to 0.2 mL min^{-1}. The injection volume was set up to 5 µL. The LC setting for other methods (LC 3, 4, 5, 6, 7, 8) can be found in SIF, Excel sheet **Table S3**.

The operating parameters of the electrospray ionization interface (ESI) are for PI mode: capillary voltage, 2500 V; end plate offset, 500 V; nebulizer, 2 bar; drying gas, 8 L min$^{−1}$; dry temperature, 200 °C; and for NI mode: capillary voltage, 3500 V; end plate offset, 500 V; nebulizer, 2 bar; drying gas, 8 L min$^{−1}$; dry temperature, 200 °C.

A QToF external calibration was performed daily with a sodium formate solution, and a segment (0.1−0.25 min) in every chromatogram was used for internal calibration, using a calibrant injection at the beginning of each run. The sodium formate calibration mixture consists of 10 mM sodium formate in a mixture of water:isopropanol (1:1). The theoretical exact masses of calibration ions in the range of 50−1000 Da were used for calibration. The instrument provided a typical resolving power of 36000−40000 during calibration (39274 at m/z 226.1593, 36923 at m/z 430.9137, and 36274 at m/z 702.8636).

SI 1.2. Deep learning deconvolutional neural network models for pH and RI

Neural Networks is comprised of layers and nodes, input and output in their architecture. These nodes are connected in each layer through a weighting function that is applied on the pervious input layer. Each layer also includes few nodes in which a function is applied to it (weighted sum of the inputs). When the number of layers between input and final output exceeding 1 layer, it is called Multiple Layer Perceptron (MLP) or Deep Learning (DL) Neural Network. In case of image-based DL method, there is a need for few extra procedures in order to learn from features (available in the image). This type of neural network is called Convolutional Neural Network (CNN). In other words, the method learns from the various part of the image and assigns weights and biases on the features map, and according to a fitting function, it can later recognize or classify the objects in the image into their defined classes. CNN is type of DL method as it requires connectivity between different nodes and layers to learn from thousands of features (pixels) extracted from an image in order to boost its classification/regression performance. The processing of extracting and then learning from features are controlled by several parameters such as Kernel size, filter values, stride, hidden layers, activation functions and number of convolutional layers. For instance, filter parameter controls the size of segments of the images created from the initial input image. The number of pixels that the filter moves every time, as it operates on the receptive field (Kernel size) of an image, is called "stride". It is important to use stride in order to enable the convolution operation to scan the whole image and not just a part of it. For example, if the size of the receptive field is 7x7, the size of the filter is 3x3 and the stride is 1, then the output of the correlation operation in the 7x7 receptive field will be a 5x5 area. By increasing "Stride" the output volume shrinks and the computational load is decreased, but if stride is too high then useful information on the image would be lost. In general, the aim is to preserve the maximum amount of information, including the low-level features contained in the input image, especially in the first layers of the network. This is due to the fact that as convolution operations are being applied, the size of the arrays can be shrieked even more and if this occurs early in the first layers of the DL architecture, then there is a risk of information loss. The first set of connected layers ends with a final layer that has no hidden layers and simply applies the "tanh" or "relu" activation function for further detection of non-linearity. These convolutional layers are connected at each node using a fully connected layer which is a deep network with n number of hidden nodes and again uses an activation function in every node. Usually, the connective layers between different convolution layers contain large number of hidden nodes while the last connected layer has hidden nodes as many as number of classes (here number of pH level which are two levels (level 1 (tR shift < 30 sec) and level 2(tR shift > 30 sec))), and 1 output node for retention index modelling used during training. It is worth saying that each layer has also one pooling layer that pools maximum areas of the previous layer to decrease the computational load required to process the data (by finding the dominant features) and also to decrease the noisy activations. Finally, the output of the whole DL is a network that applies the "softmax" or linear output activation functions for classification and regression case, respectively. "softmax" is suitable for classification problems because its output can be interpreted as a probability distribution over all possible classes. The following structures were followed to construct DL-CNN models:

To model pH effect:

- First convolutional layer:
- Convolution layer includes; 5×5 Kernel size, 20 filters, "tanh" activation, max pool size 2 × 2, stride 2×2, dropout probability of 0.3
- Second convolutional layer:
- Convolution layer includes; 5×5 Kernel size, 30 filters, "tanh" activation, max pool size 2 × 2, stride 2×2, dropout probability of 0.3
- First fully connected layer:
- Hidden nodes of 850, "tanh" activation and dropout probability of 0.3
- Second fully connected layer:
- Hidden nodes of 2 (as of two classes pH uncertainty)
- Output layer
- Softmax function
- Fitness function: miss-classification error (mx.accuracy)

To model Retention Index values:

- First convolutional layer:
- Convolution layer includes; 9×9 Kernel size, 10 filters, “tanh” activation, max pool size 2 × 2, stride 2×2, dropout probability of 0.1
- Fully connected layer:
- Hidden nodes of 60, “relu” activation and dropout probability of 0.1
- Hidden nodes of 1
- Output layer (linear regression output)
- Fitness function: RMSE (root mean square error)

SI 2. Results and Discussion

SI 2.1. Error distribution of RI models

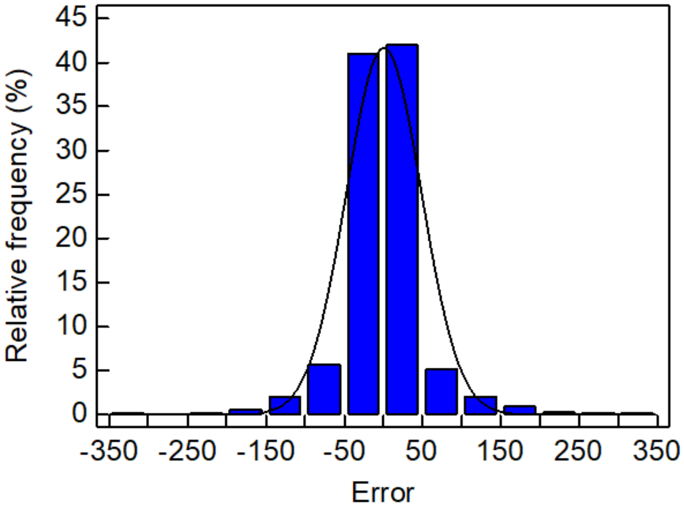
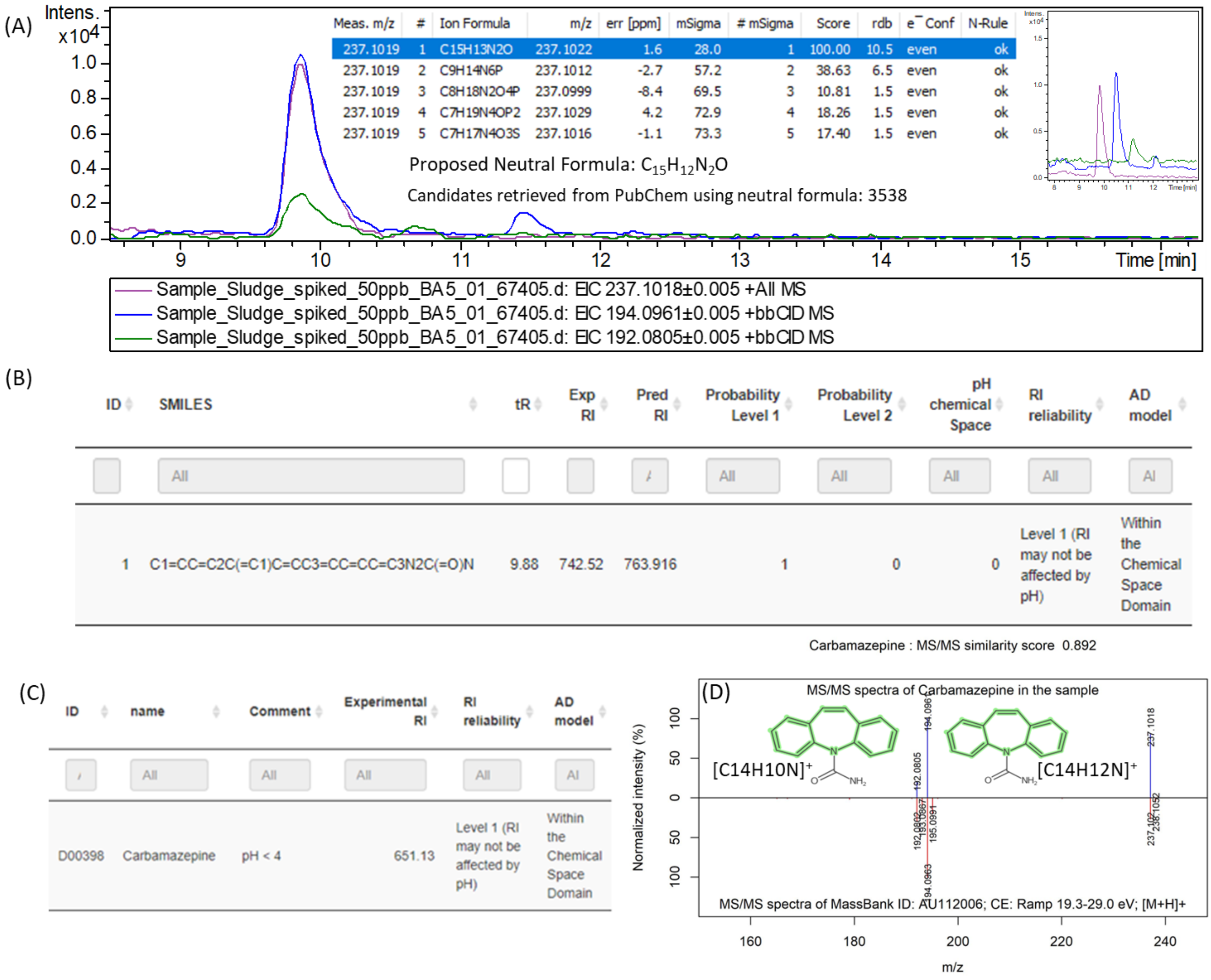


Figure S1. Distribution of error for CDEA RI models (both pH<4 and pH>6)

SI 2.2. Application of CDEA-RI in sludge sample



Meas. m/z	#	Ion Formula	m/z	err [ppm]	mSigma	# mSigma	Score	rdb	e⁻ Conf	N-Rule
237.1019	1	C15H13N2O	237.1022	1.6	28.0	1	100.00	10.5	even	ok
237.1019	2	C9H14N6P	237.1012	-2.7	57.2	2	38.63	6.5	even	ok
237.1019	3	C8H18N2O4P	237.0999	-8.4	69.5	3	10.81	1.5	even	ok
237.1019	4	C7H19N4OP2	237.1029	4.2	72.9	4	18.26	1.5	even	ok
237.1019	5	C7H17N4O3S	237.1016	-1.1	73.3	5	17.40	1.5	even	ok

Proposed Neutral Formula: $C_{15}H_{12}N_2O$

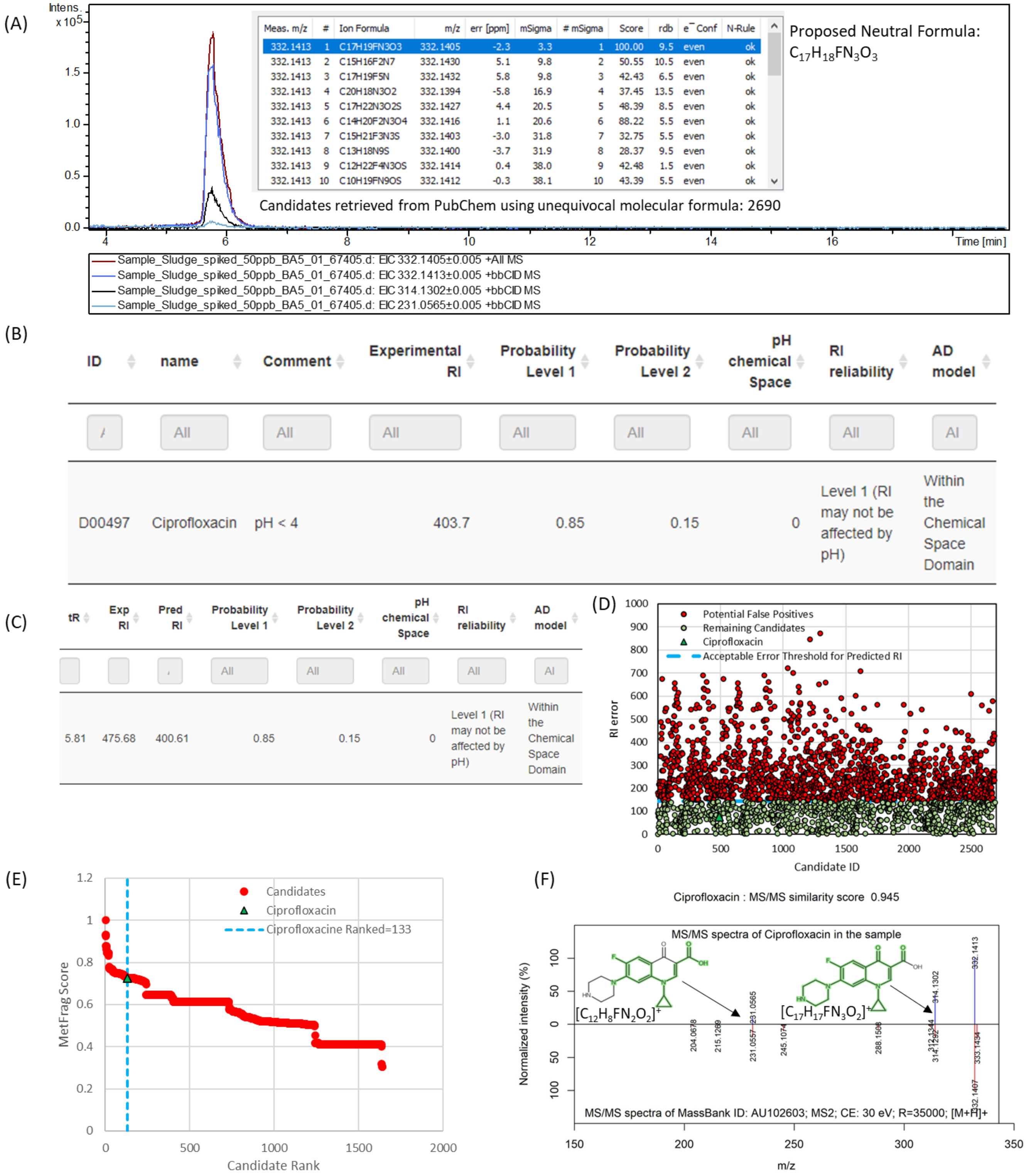
Candidates retrieved from PubChem using neutral formula: 3538

Sample_Sludge_spiked_50ppb_BA5_01_67405.d: EIC 237.1018±0.005 +All MS
Sample_Sludge_spiked_50ppb_BA5_01_67405.d: EIC 194.0961±0.005 +bbCID MS
Sample_Sludge_spiked_50ppb_BA5_01_67405.d: EIC 192.0805±0.005 +bbCID MS

ID	SMILES	tR	Exp RI	Pred RI	Probability Level 1	Probability Level 2	pH chemical Space	RI reliability	AD model
	All				All	All	All	All	Al
1	C1=CC=C2C(=C1)C=CC3=CC=CC=C3N2C(=O)N	9.88	742.52	763.916	1	0	0	Level 1 (RI may not be affected by pH)	Within the Chemical Space Domain

ID	name	Comment	Experimental RI	RI reliability	AD model
	All	All	All	All	Al
D00398	Carbamazepine	pH < 4	651.13	Level 1 (RI may not be affected by pH)	Within the Chemical Space Domain

Figure S2. Identification case C01: (A) Detection of m/z 237.1018, related DIA fragments and assigned formula in sludge sample analyzed by LC 8 setting; B) Removal of False positives by CDEA RI and retaining the best predicted and experimental RI match; C) RI bank hit for m/z 237.1019 and matching RI values between library and RI value from sludge samples analyzed in LC 8 under 232 error (RI unit) window; D) Matching MS/MS fragments with MassBank record and annotated substructures



Meas. m/z	#	Ion Formula	m/z	err [ppm]	mSigma	# mSigma	Score	rdb	e¯ Conf	N-Rule
332.1413	1	C17H19FN3O3	332.1405	-2.3	3.3	1	100.00	9.5	even	ok
332.1413	2	C15H16F2N7	332.1430	5.1	9.8	2	50.55	10.5	even	ok
332.1413	3	C17H19F5N	332.1432	5.8	9.8	3	42.43	6.5	even	ok
332.1413	4	C20H18N3O2	332.1394	-5.8	16.9	4	37.45	13.5	even	ok
332.1413	5	C17H22N3O2S	332.1427	4.4	20.5	5	48.39	8.5	even	ok
332.1413	6	C14H20F2N3O4	332.1416	1.1	20.6	6	88.22	5.5	even	ok
332.1413	7	C15H21F3N3S	332.1403	-3.0	31.8	7	32.75	5.5	even	ok
332.1413	8	C13H18N9S	332.1400	-3.7	31.9	8	28.37	9.5	even	ok
332.1413	9	C12H22F4N3OS	332.1414	0.4	38.0	9	42.48	1.5	even	ok
332.1413	10	C10H19FN9OS	332.1412	-0.3	38.1	10	43.39	5.5	even	ok

Proposed Neutral Formula: $C_{17}H_{18}FN_3O_3$

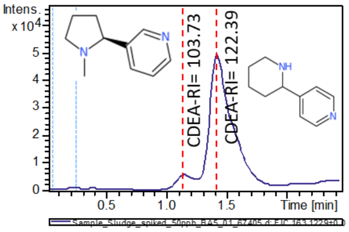
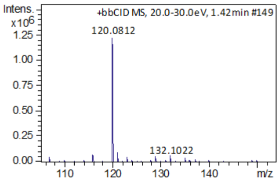
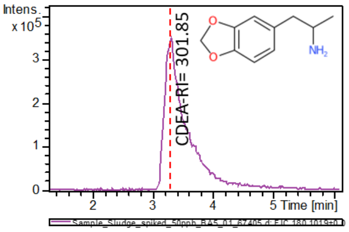
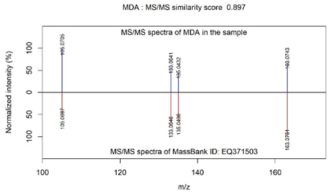
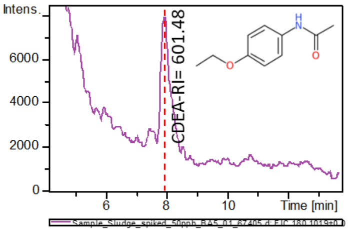
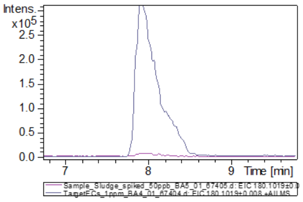
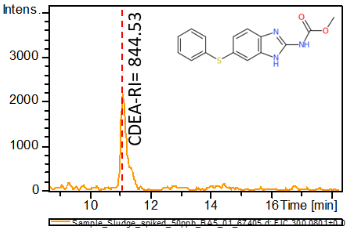
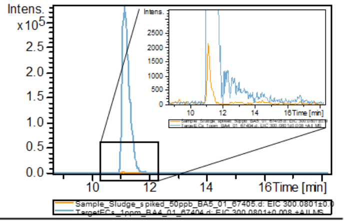
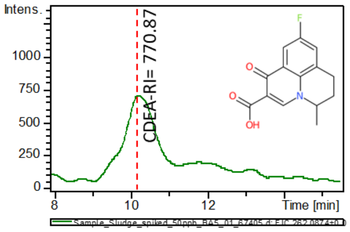
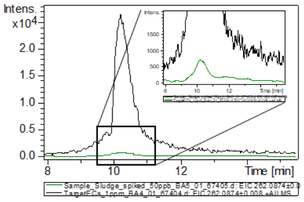
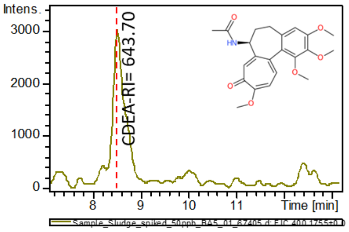
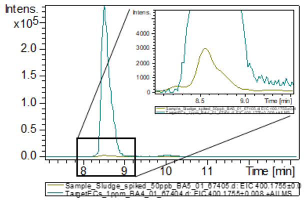
Candidates retrieved from PubChem using unequivocal molecular formula: 2690

ID	name	Comment	Experimental RI	Probability Level 1	Probability Level 2	pH chemical Space	RI reliability	AD model
/	All	All	All	All	All	All	All	Al
D00497	Ciprofloxacin	pH < 4	403.7	0.85	0.15	0	Level 1 (RI may not be affected by pH)	Within the Chemical Space Domain

tR	Exp RI	Pred RI	Probability Level 1	Probability Level 2	pH chemical Space	RI reliability	AD model
		,	All	All	All	All	Al
5.81	475.68	400.61	0.85	0.15	0	Level 1 (RI may not be affected by pH)	Within the Chemical Space Domain

(A)
Intens.
x 10^5
1.5
1.0
0.5
0.0
4
6
8
10
12
14
16
Time [min]
Sample_Sludge_spiked_50ppb_BA5_01_67405.d: EIC 332.1405±0.005 +All MS
Sample_Sludge_spiked_50ppb_BA5_01_67405.d: EIC 332.1413±0.005 +bbCID MS
Sample_Sludge_spiked_50ppb_BA5_01_67405.d: EIC 314.1302±0.005 +bbCID MS
Sample_Sludge_spiked_50ppb_BA5_01_67405.d: EIC 231.0565±0.005 +bbCID MS
(B)
(C)
(D)
Potential False Positives
Remaining Candidates
Ciprofloxacin
Acceptable Error Threshold for Predicted RI
RI error
1000
900
800
700
600
500
400
300
200
100
0
0
500
1000
1500
2000
2500
Candidate ID
(E)
Candidates
Ciprofloxacin
Ciprofloxacine Ranked=133
MetFrag Score
1.2
1
0.8
0.6
0.4
0.2
0
0
500
1000
1500
2000
Candidate Rank
(F)
Ciprofloxacin : MS/MS similarity score 0.945
MS/MS spectra of Ciprofloxacin in the sample
Normalized intensity (%)
100
50
0
50
100
[C12H8FN2O2]+
[C17H17FN3O2]+
F
O
OH
N
HN
204.0678
215.1269
231.0565
231.0557
245.1074
288.1508
312.1344
314.1302
314.1292
332.1413
333.1434
332.1407
MS/MS spectra of MassBank ID: AU102603; MS2; CE: 30 eV; R=35000; [M+H]+
150
200
250
300
350
m/z

Figure S3. Identification case C02: (A) Detection of m/z 332.1405, its fragments based on DIA mode and assigned formula in sludge sample analyzed by LC 8 setting; B) RI bank hit for m/z 332.1405 and matching RI values between library and RI value from sludge samples analyzed in LC 8; C) Predicted RI values for Ciprofloxacin and associated uncertainty; D) Removal of False positives by CDEA RI and retaining the best predicted and experimental RI match; E) Candidates ranked by MetFrag (1041 were excluded after the process due to not explaining any input fragments); F) Matching MS/MS fragments with MassBank record

ID	Extracted Ion Chromatogram (EIC)	m/z [M+H]+	pH effect (probabilities) Class 1: Class 2	Experimental RI from CDEA RI bank	Experimental RI in sludge sample	Note
C03		163.1224	Nicotine: 1.000 : 0.000 Anabasine: 0.806 : 0.194	Nicotine: 191.07 Anabasine: 248.86	103.73 122.39	Diagnostic fragments for Anabasine: 120.0808 (MassBank ID: NA000940) is observed: 
C04		180.1013	MDA: 0.017 : 0.983	MDA: 380.81 (pH=3.6)	301.85 (pH=2.6)	Top 4 fragments from MassBank ID: EQ371503: 
C05		180.1013	Phenacetin: 0.997 : 0.003	Phenacetin: 538.52	601.48	It is detected at low intensity in sludge sample, no MS/MS fragments are obtained; verified by reference standard: 
C06		300.0801	Fenbendazole: 0.977 : 0.023	Fenbendazole: 917.07	844.53	It is detected at low intensity in sludge sample, no MS/MS fragments are obtained; verified by reference standard: 
C07		262.0874	Flumequine: 0.001 : 0.999	Flumequine: 658.65 (pH=3.6)	770.87 (pH=2.6)	It is detected at low intensity in sludge sample, no MS/MS fragments are obtained; verified by reference standard: 
C08		400.1755	Colchicine: 0.997 : 0.003	Colchicine: 579.51	643.70	It is detected at low intensity in sludge sample, no MS/MS fragments are obtained; verified by reference standard: 

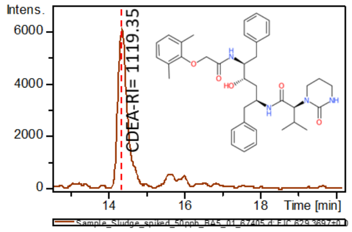
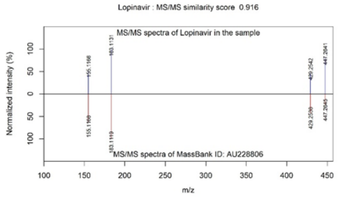
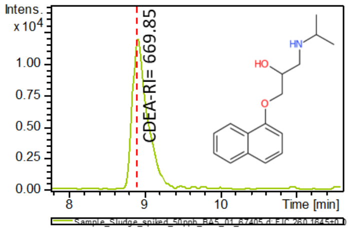
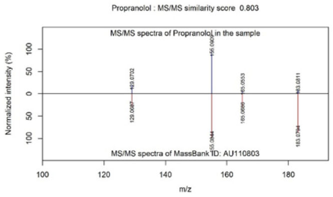
ID	Extracted Ion Chromatogram (EIC)	m/z [M+H]+	pH effect (probabilities) Class 1: Class 2	Experimental RI from CDEA RI bank	Experimental RI in sludge sample	Note
C09		629.3697	Lopinavir: 0.312 : 0.688	Lopinavir: 1109.23 (pH=3.6)	1119.35 (pH=2.6)	Top 4 fragments from MassBank ID: AU228806: 
C10		260.1645	Propranolol: 0.999 : 0.001	Propranolol: 592.62	669.85	Top 4 fragments from MassBank ID: AU110803: 

Figure S4. Extended application of CDEA RI bank in sludge samples

SI 2.3. Suspect screening of nitrosamines

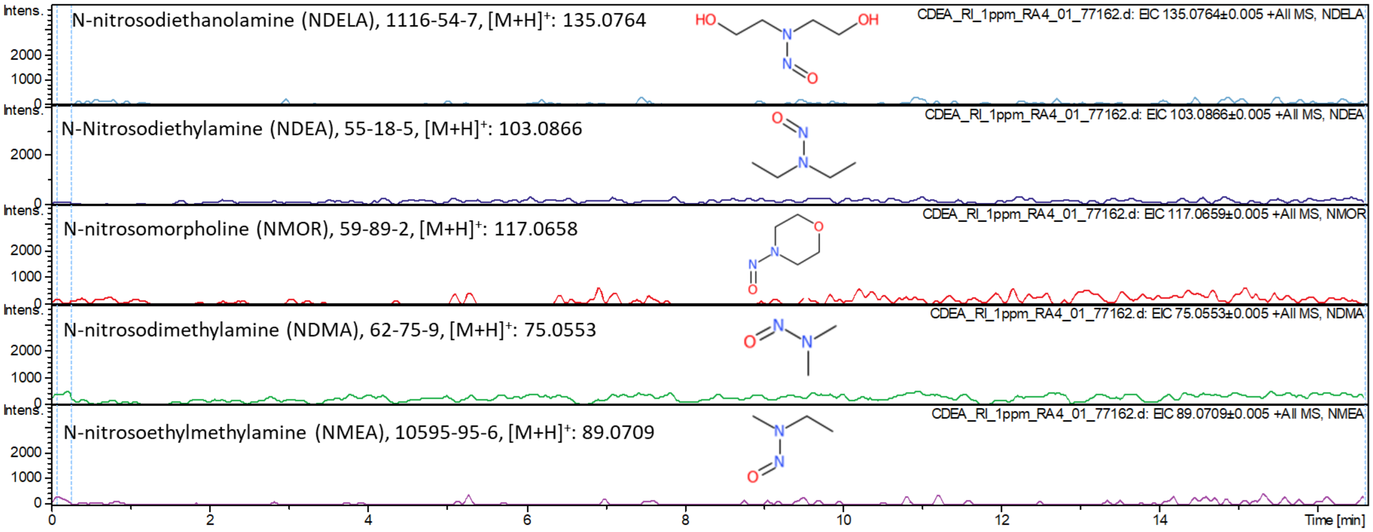


Figure S5. EIC of 5 most common nitrosamines reported in the literature in products which contain DEA; none of these impurities are detected in the synthesized CDEAs

SI 2.4. Stability Test

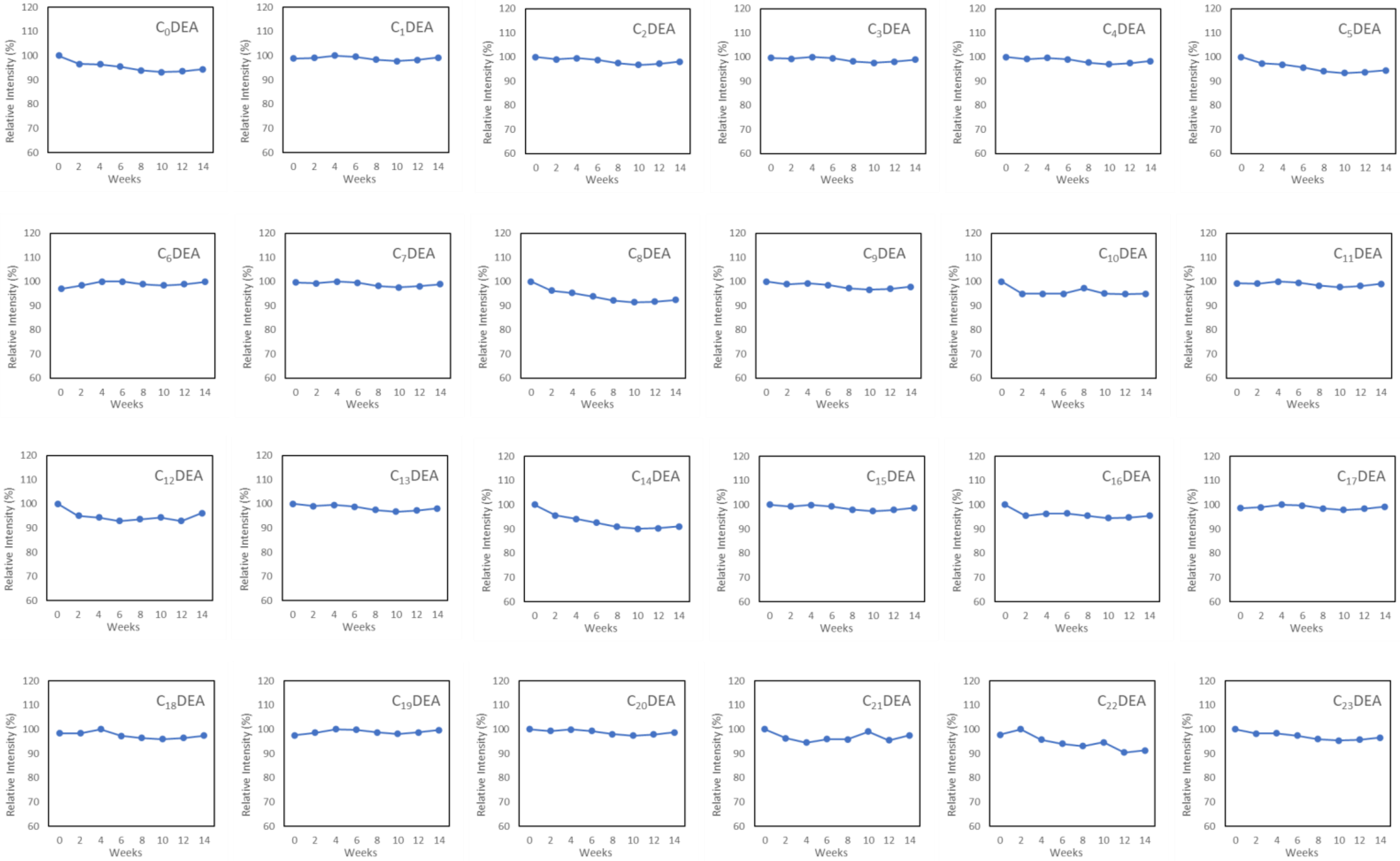


Figure S6. Three-month stability test of CDEA RIs stored in 8°C